DISCRETE DIFFUSION FOR SINGLE-CELL GENE EXPRESSION MODELING

Sanjukta Bhattacharya[1], **Christian Gensbigler**[2], **Shaamil Karim**[2], **Jon Lees**[1]
[1] University of Bristol [2] Dartmouth College
{ae25872, jon.lees}@bristol.ac.uk
{christian.a.gensbigler.28, shaamil.k.shaw.alem.26}@dartmouth.edu

ABSTRACT

Current generative modeling of single-cell transcriptomics relies on continuous latent representations, transforming inherently discrete and sparse gene counts into continuous space. We propose Discrete Cell Models (DCM), a diffusion-based framework that learns cellular representations directly in the discrete domain. Our framework supports both unconditional and conditional generation, allowing for precise modeling of complex biological scenarios such as cell-type-specific transcriptional responses to genetic perturbations. We demonstrate that DCM scales effectively and achieves strong performance against current state-of-the-art methods, including scVI, CPA, STATE, scGPT, and scLDM. On the Dentate Gyrus benchmark, DCM achieves a 5-fold improvement in MMD^2RBF and a nearly 2-fold improvement in W_2 distance, over the leading continuous diffusion baseline (scLDM). On the conditional Replogle perturbation benchmark, DCM sets a new state of the art on W_2 distance while remaining competitive on MMD^2RBF. Together, these results establish discrete diffusion as a promising direction for foundational models of cellular biology.

1 INTRODUCTION

A cell can be viewed as an information-processing system that maps environmental inputs to molecular responses through an internal regulatory program Fitch (2021). While the structure of the regulatory network itself is not directly accessible, single-cell RNA sequencing provides a discrete high-dimensional but partial observation of its activity Macosko et al. (2015); Gulati et al. (2020); Zeng & Dai (2019). Crucially, this observation is discrete: we do not measure continuous expression levels, but count individual molecular events. Across many cells and conditions, these observations implicitly encode the statistical dependencies induced by the underlying regulatory structure. The objective of a virtual cell model is to learn these dependencies well enough to generate realistic cellular states and to predict how modifications to the regulatory program, such as genetic perturbations, alter the resulting transcriptome Dixit et al. (2016); Adamson et al. (2016).

The central object of these observations is the count matrix: a sparse, integer-valued tensor recording how many mRNA molecules of each gene were captured from each cell. Yet the dominant generative models have been applied to this problem begin by embedding these integers into continuous vector spaces Lopez et al. (2018b); Gandhi et al. (2025); Bereket & Karaletsos (2023); Adduri et al. (2025). Variational autoencoders such as scVI Lopez et al. (2018a); Kingma & Welling (2013) learn low-dimensional latent representations under a negative binomial observation model, accounting for the count nature of the data at the likelihood level but restricting the latent prior to a unimodal Gaussian. Compositional Perturbation Autoencoder (CPA) Lotfollahi et al. (2023) extends this framework by disentangling cell identity from perturbation effects in latent space, enabling combinatorial prediction of unseen conditions. Transformer-based foundation models like scGPT Cui et al. (2024) scale to tens of millions of cells and produce powerful embeddings. Although operating in the discrete space, scGPT's autoregressive generation imposes a sequential ordering on genes, which is an inductive bias that contrasts with the exchangeability of gene expression Kim et al. (2025).

More recently, diffusion-based approaches have addressed generation more directly. scDiffusion Luo et al. (2024) applies classifier-guided continuous diffusion in the latent space of a pre-trained

foundation model. CFGen Palma et al. (2024) introduces flow matching Lipman et al. (2022) with compositional guidance and explicit negative binomial likelihoods, enabling generation conditioned on multiple biological attributes simultaneously. scLDM Palla et al. (2025), the current state of the art, combines a fully transformer-based VAE, with permutation-invariant encoding and equivariant decoding, with a latent diffusion model parameterized by Diffusion Transformers Peebles & Xie (2023) and trained via flow matching. scLDM demonstrated that architectural respect for the exchangeability of genes yields substantial gains in generation fidelity. Yet despite their differences in architecture and training objective, all of these recent methods share a common representational commitment: they transform raw integer counts into continuous vectors and model in the resulting real-valued space. The discrete structure of the raw measurements is recovered only after sampling, through rounding or sampling from count distributions fitted to continuous predictions.

This representational choice warrants scrutiny. In language modeling, we do not embed discrete tokens into continuous space, generate in that space, then discretize back—we model tokens directly Vaswani et al. (2017). Yet for transcript counts, the field has adopted continuous relaxation as a default Lopez et al. (2018b); Gandhi et al. (2025); Bereket & Karaletsos (2023); Adduri et al. (2025). The standard justification may be that continuous spaces are easier to model, and that the observation model (Poisson, negative binomial) accounts for discreteness at the likelihood level. We argue that this justification is insufficient, and that continuous relaxation introduces fundamental representational limitations.

First, continuous models assign probability mass to non-integer values that correspond to no possible measurement—model capacity spent on impossible states. Second, the natural metric on count space is not Euclidean: the difference between 0 and 1 transcript (presence versus absence of expression) is biologically distinct from the difference between 100 and 101 transcripts (sampling noise) Hafemeister & Satija (2019). Continuous embeddings with Euclidean metrics cannot represent this asymmetry without learning it from data; discrete models respect it by construction. Third, gene regulatory networks induce relations that depend on the functional presence of gene products, which for lowly-expressed genes is inherently stochastic and discrete—a gene with mean expression 0.5 transcripts/cell is "off" in most cells and "on" in some, a bimodal phenomenon that continuous models must learn to represent but that discrete models capture naturally Elowitz et al. (2002). Finally, there is an information-theoretic argument: when the true data-generating process is discrete, any continuous relaxation introduces a representation gap, forcing the model to learn discretization boundaries rather than structure within the discrete space Maddison et al. (2016). Furthermore, recent advances in discrete diffusion for text Lou et al. (2023); Sahoo et al. (2024) and protein sequences Tang et al. (2024); Gruver et al. (2023) have demonstrated that operating directly in discrete token space yields substantial improvements over continuous relaxations—results that suggest similar gains should be achievable for the discrete, sparse count distributions characteristic of single-cell transcriptomics.

We introduce Discrete Cell Models (DCM), a framework that applies Score Entropy Discrete Diffusion (SEDD) Lou et al. (2023); Campbell et al. (2022) directly to raw transcript counts, eliminating the continuous relaxation used by prior work. DCM treats each gene's expression level as a discrete token and learns to reverse a corruption process over the count space via the concrete score: the ratio of data distribution across neighbouring discrete states Meng et al. (2022). The model supports multi-conditional and unconditional generation within a single end-to-end architecture, synthesizing cell states conditioned jointly on cell type, perturbation identity, and other biological attributes. We evaluate DCM on two complementary benchmarks. On the Dentate Gyrus dataset (unconditional generation)Hochgerner et al. (2018), DCM achieves a nearly 2-fold reduction in W_2 distance and a 5-fold improvement in MMD^2RBF over scLDM. On the Replogle dataset (conditional perturbation prediction), DCM achieves the best W_2 distance across all baselines while maintaing a competitive MMD^2RBF. Together, these results establish discrete diffusion as a promising direction for foundational models of cellular biology.

2 Discrete diffusion for gene expression generation

Gene Expression Representation. We represent single-cell gene expression profiles as discrete sequences. Let $\mathbf{x} \in \mathcal{X}^M$ denote a cell's expression profile across M genes, where $\mathcal{X} = \{0, 1, \ldots, K\}$ is the discrete vocabulary of binned or raw expression counts, with K as the max gene expression count in the dataset. Each element $x_i \in \mathcal{X}$ represents the expression level of gene i in a single cell.

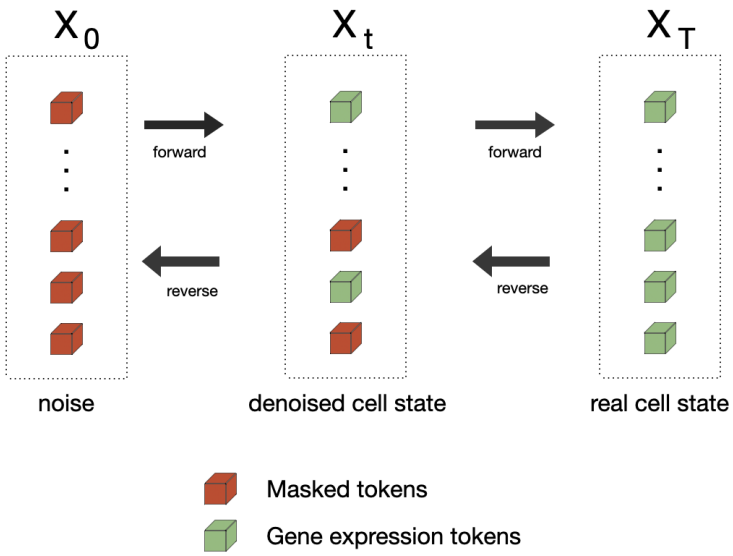

Figure 1: Generating single cell data with DCM

Conditional Generation Objective. We aim to learn a generative model $p_\theta(\mathbf{x} \mid \mathbf{c})$ that produces gene expression profiles conditioned on cellular attributes $\mathbf{c} = (\text{pert}, \text{cell\_type})$, where pert denotes the genetic perturbation (e.g., gene knockdown) and cell_type specifies the cellular context (e.g., HepG2, Jurkat).

Given observed i.i.d. data $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{c}_n)\}_{n=1}^N$, where each observation consists of a gene expression profile $\mathbf{x}_n$ and its corresponding conditions $\mathbf{c}_n$, our objective is to learn a parametric model $p_\theta(\mathbf{x} \mid \mathbf{c})$ with parameters θ that maximizes the conditional log-likelihood:

$$\ell(\theta; \mathcal{D}) = \sum_{n=1}^{N} \log p_\theta(\mathbf{x}_n \mid \mathbf{c}_n) \tag{1}$$

Discrete Diffusion Framework. We use Score Entropy Discrete Diffusion (SEDD) Lou et al. (2023), which models the data distribution through a continuous-time discrete-state markov process. Rather than directly parameterizing $p_\theta(\mathbf{x} \mid \mathbf{c})$, SEDD learns to estimate ratios of the data distribution at different noise levels which is then used to sample the gene expressions.

Forward Diffusion Process. We define a continuous-time markov process that progressively corrupts clean expression profiles $\mathbf{x}^0 \sim p_{\text{data}}(\mathbf{x} \mid \mathbf{c})$ through independent token-level transitions. For time $t \in [0, 1]$, the forward process evolves according to:

$$\frac{dp_t}{dt} = Q_t p_t, \quad p_0 \approx p_{\text{data}}(\cdot \mid \mathbf{c}) \tag{2}$$

where $Q_t \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ is the token-level diffusion matrix. Each element $Q(i, j)$ of the matrix Q represents the probability that the token at position i transitions into the token value at position j which defined over the vocabulary. We use an *absorbing* diffusion structure where all tokens transition toward a special 'MASK' state. We use a log-linear noise schedule $Q_t = \sigma(t) Q^{\text{absorb}}$ where $\sigma(t) = \frac{1-\varepsilon}{\varepsilon + (1-\varepsilon)t}$ and $\varepsilon = 10^{-3}$.

Reverse Process via Concrete Scores. The reverse diffusion process is characterized by the time-reversed matrix $Q_t^\dagger$ with entries:

$$Q_t^\dagger(y, x) = \frac{p_t(y)}{p_t(x)} Q_t(x, y), \quad x \neq y \tag{3}$$

The ratios $\frac{p_t(y)}{p_t(x)}$ are called *concrete scores* [Meng et al. (2022)], the discrete analog of $\nabla_x \log p_t(x)$ in continuous diffusion. For gene expression sequences, we parameterize a score network s_θ :

$\mathcal{X}^M \times [0,1] \to \mathbb{R}^{M \times |\mathcal{X}|}$ that estimates ratios between sequences with Hamming distance 1:

$$s_\theta(\mathbf{x}^t, t)_{i,v} \approx \frac{p_t(x_1^t \dots v \dots x_M^t)}{p_t(x_1^t \dots x_i^t \dots x_M^t)} \tag{4}$$

where position i has token x_i^t replaced by token v. Conditioning on $\mathbf{c}$ is incorporated by augmenting the network input: $s_\theta(\mathbf{x}^t, t, \mathbf{c})$.

Training Objective: Denoising Cross-Entropy. Following recent theoretical analysis Ou et al. (2025), for the absorbing case, the DWDSE objective simplifies to a weighted cross-entropy loss. Specifically, the concrete score can be reparameterized as:

$$s_\theta(\mathbf{x}^t, t, \mathbf{c})_{i,v} = \frac{\alpha(t)}{1-\alpha(t)} \cdot p_\theta(x_i^0 = v \mid \mathbf{x}^t, \mathbf{c}) \tag{5}$$

where $\alpha(t) = e^{-\int_0^t \sigma(s)ds}$ is the survival probability at time t. Under this parameterization, the DWDSE loss reduces to:

$$\mathcal{L} = \mathbb{E}_{t\sim\text{Uniform}(0,1),\mathbf{x}^0\sim p_{\text{data}}(\cdot|\mathbf{c}),\mathbf{x}^t\sim q(\cdot|\mathbf{x}^0)} \left[\sum_{i=1}^{M} \text{CrossEntropy}\left(p_\theta(x_i^0 \mid \mathbf{x}^t, t, \mathbf{c}), x_i^0\right) \right] \tag{6}$$

This formulation enables tractable likelihood-based training while naturally handling the discrete, high-dimensional nature of gene expression data.

3 EXPERIMENTS

In this section, we present evaluations that access the effectiveness of proposed DCM-framework to learn and infer cell state representations. To evaluate the quality of generated gene expression profiles, we compare the distribution of predicted gene expression vectors to that of ground-truth expression vectors using two complementary metrics: Maximum Mean Discrepancy (MMD) and the 2-Wasserstein distance (W_2). Together, these metrics assess both fine-grained statistical similarity and global geometric alignment between distributions. We use these two metrics to provide a complementary view of model performance. We benchmark our model against SCVI [Lopez et al. (2018a)], scDiffusion [Luo et al. (2024)], CFGen [Palma et al. (2024)], and the current SOTA scDLM [Palla et al. (2025)], with model scores taken from [Palla et al. (2025)].

The **Maximum Mean Discrepancy (**MMD**)** is a non-parametric metric that measures the distance between two distributions by comparing their expectations under a rich class of functions. In our setting, MMD quantifies how well the distribution of predicted gene expression vectors matches the distribution of true gene expression vectors. Given ground-truth samples $X = \{x_1, x_2, \cdots, x_m\}$ and predicted samples $Y = \{y_1, y_2, \cdots, y_m\}$, we compute the unbiased empirical estimator:

$$\widehat{\text{MMD}}^2(X,Y) = \frac{1}{m(m-1)} \sum_{i\neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i\neq j} k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j) \tag{7}$$

and we use a Gaussian RBF kernel here: $k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$ which makes MMD sensitive to differences in higher-order statistics and multimodal structure.

The **2-Wasserstein distance** measures the discrepancy between distributions based on optimal transport and directly captures the geometry of the gene expression space. Intuitively, it quantifies the minimum cost of transporting mass from the predicted distribution to the ground-truth distribution. For empirical distributions with equal sample size n, this reduces to a discrete matching problem:

$$W_2^2(X,Y) = \frac{1}{n} \min_{\pi\in\Pi_n} \sum_{i=1}^{n} \|x_i - y_{\pi(i)}\|^2 \tag{8}$$

where Π_n is the set of all permutations of $\{1, \cdots, n\}$. When both predicted and true gene expression distributions are approximated as Gaussians with means μ_P, μ_Q and covariances $\sum_P, \sum_Q$, we use the closed-form expression:

$$W_2^2(P,Q) = \|\mu_P - \mu_Q\|^2 + \text{tr}\left(\Sigma_P + \Sigma_Q - 2(\Sigma_P^{1/2}\Sigma_Q\Sigma_P^{1/2})^{1/2}\right) \tag{9}$$

which coincides with the Fréchet Distance (FD). In our evaluation, a lower W_2 indicates that predicted gene expression vectors align closely with ground-truth vectors in terms of global location and covariance structure. We note that some prior works report metrics computed in PCA-reduced space Palma et al. (2024). Since DCM operates directly on raw gene expression counts without dimensionality reduction, we evaluate all metrics in the full gene expression space for consistency.

Our experiments involve two major directions of evaluations: unconditional and conditional generative performances on both observational and perturbational datasets. We employ a DiT backbone Peebles & Xie (2023) for the score network where the diffusion-time condtioning is done via Adaptive LayerNorm (AdaLN). Here we introduce other conditioning variables to the score network as we later describe. We use full gene expression profiles for a cell type as the context length of the transformer ($\approx 17k$) and employ Flash Attention Dao et al. (2022) for efficient attention computation. To handle variable gene subsets across cells: we use special 'PAD' tokens for non-expressed or non-selected genes and mask attention to these tokens to prevent them from influencing the model's computations. To reduce training time, we use lower precision representations of the gene expression.

3.1 Unconditional generation

For this experiment, we are interested in evaluating the model's fundamental ability to learn and reproduce the underlying distribution of single-cell gene expression data without guidance from conditioning labels with the only conditioning to the score network being the diffusion-time embeddings. Here, we use single cell RNA-sequencing data from the datasets used in both [Palla et al. (2025)], and [Palma et al. (2024)]. We evaluate the model's output to match the distribution of expression vectors to real cell states that were held out during the training of the model.

Table 1: Model generation performance on unconditional cell generation benchmarks, for the dentate gyrus dataset.

Dataset	Model	W2 $\downarrow$	MMD^2 RBF $\downarrow$
Dentate Gyrus	scDiffusion	17.321 ± 0.041	0.689 ± 0.000
	CFGen	11.608 ± 0.066	0.075 ± 0.000
	scLDM	10.615 ± 0.028	0.102 ± 0.003
	DCM (5M)	$\mathbf{5.913 \pm 0.091}$	$\mathbf{0.019 \pm 0.005}$

3.1.1 Results & discussion

DCM achieves substantial improvements over all baselines on both metrics. On $W2$ distance, DCM (5.913) reduces the gap to the ground truth nearly 2-fold compared to scLDM (10.615), the strongest continuous baseline. On MMD^2RBF, DCM (0.019) achieves a 5-fold improvement over the closest baseline CFGen (0.075) and a further improvement over scLDM (0.102).

We offer two candidate explanations for these gains, which remain speculative without further ablation. First, single-cell data is zero-inflated, and discrete models represent zeros as a distinct state rather than requiring a sharp mode at zero in continuous space. Second, for lowly-expressed genes, neighboring count values (0, 1, 2) carry categorical significance that continuous interpolation may obscure. Whether DCM's gains are concentrated in these regimes is an empirical question we leave to future work.

DCM also achieves these results with a 5M parameter model, which is substantially smaller than scLDM's two-stage architecture (transformer VAE + diffusion transformer). The end-to-end discrete diffusion framework eliminates the need for a separate encoder-decoder pipeline, simplifying training and reducing the number of design choices.

3.2 Conditional generation

In this experiment, we train our model to generate gene expression conditioned on multiple attributes: cell type or perturbation type (gene knockouts) or both. For cell types, we create conditioning embedding via one-hot encoding of a specific cell type followed by a linear layer. For perturbation type, we employ a protein language model Lin et al. (2023) to obtain the embeddings

for each of the perturbation labels in the dataset, enabling generalisation on held-out labels. The conditioning embeddings are then concatenated with the diffusion-time embeddings and provided to the score network via the AdaLN mechanism. At inference time, DCM is queried with combinations of cell type and perturbations to generate new gene expression profiles.

Table 2: Model generation performance on conditional perturbation prediction benchmarks (Replogle dataset). DCM is evaluated at two scales: 20M parameters on the full dataset (4 cell lines) and 10M parameters on K562 only. The K562-only setting isolates perturbation modeling from cross-cell-line variation.

Dataset	**Model**	**W2** $\downarrow$	**MMD^2 RBF** $\downarrow$
Replogle	scVI	17.359 ± 0.051	0.453 ± 0.003
	CPA	11.510 ± 0.029	0.532 ± 0.003
	scGPT	34.166 ± 0.272	3.087 ± 0.010
	STATE	20.58 ± 0.039	0.730 ± 0.003
	scLDM ($\omega = 1$)	11.292 ± 0.033	$\mathbf{0.200 \pm 0.002}$
	scLDM ($\omega = 5$)	12.900 ± 0.069	0.320 ± 0.004
	scLDM ($\omega = 10$)	14.911 ± 0.091	0.436 ± 0.005
	DCM (20M)	$\mathbf{10.03 \pm 0.337}$	0.688 ± 0.099
Replogle - K562	**DCM (10M)**	7.284 ± 0.091	0.605 ± 0.005

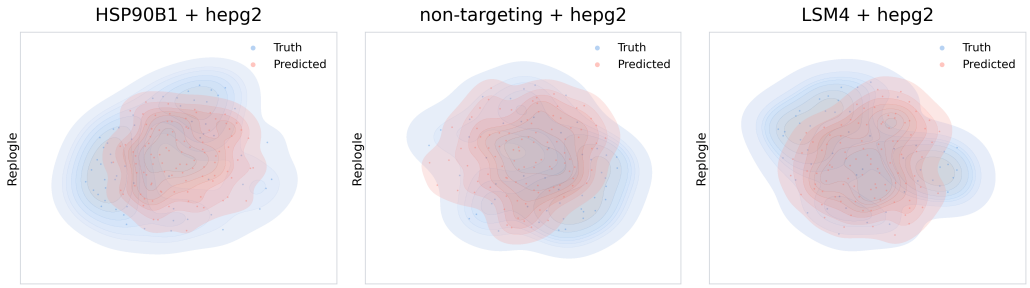

Figure 2: UMAPs from Replogle: (Left) For the HSP90B1 & (Right) LSM4 perturbation, in HepG2 cells, predicted (red) and true (blue) density contours show strong overlap, indicating that the model captures the perturbation-specific transcriptional shift. (Center) Non-targeting control in HepG2 cells confirms the model reproduces the unperturbed baseline distribution.

3.2.1 Results & discussion

In table 2, DCM achieves the best W_2 distance on the full Replogle benchmark (10.03 vs. 11.292 for scLDM at w = 1), representing an 13% improvement in global distributional alignment. On the K562-only evaluation, which isolates perturbation modeling from cross-cell-line heterogeneity in the Replogle dataset, DCM achieves a competetive W_2 of 7.284. These results demonstrate that discrete diffusion captures population-level transcriptomic structure effectively in the conditional setting.

However, we observe a divergence between W_2 and MMD^2RBF on these benchmarks: DCM's MMD^2 (0.688) is higher than scLDM's (0.200) for the Replogle benchmark. The W_2 metric (Eq. 9) captures mean and covariance alignment—the first two moments of the distribution. The MMD^2RBF metric is sensitive to higher-order distributional features, including gene-gene correlations and tail behavior. DCM's strong W_2 but weaker MMD^2RBF therefore suggests that it accurately recovers the mean expression profile and gene-level variances for each perturbation condition, but introduces errors in higher-order dependency structure.

We consider three candidate explanations for this pattern. First, our conditioning mechanism, additive embedding concatenation, may not fully capture interactions between perturbation and cell type. The same knockdown produces different effects in different cell lines, and capturing this may

require multiplicative or attention-based conditioning; scLDM's cross-attention mechanism may provide this capacity. Supporting this hypothesis, MMD^2_{RBF} improves in the K562-only setting where cell-type interactions are absent, and is strong on unconditional Dentate Gyrus where no conditioning is required. The performance gap therefore may be partly attributable to the conditioning mechanism rather than the core discrete diffusion approach itself. To disentangle these factors, future work should explore more sophisticated guidance mechanisms suited to discrete modelling such as classifier-free guidance over discrete token spaces or attention-based perturbation conditioning to more rigorously evaluate this hypothesis.

Second, gene-gene correlations induced by perturbations may be inherently better captured in continuous latent spaces, which provide smooth interpolation between related states. The MMD^2RBF gap persists even on K562 alone (0.605 vs. scLDM's 0.200), though conditioning differences confound this comparison. Third, the SEDD training objective may be less effective than flow matching for capturing correlational structure, independent of the discrete-versus-continuous distinction.

Overall, discrete diffusion achieves state-of-the-art performance on W_2 for conditional perturbation prediction, demonstrating that the approach extends effectively from unconditional to conditional generation. Ablations to isolate the source of the MMD^2RBF gap are left to future work. We also note that the sequencing-depth of the datasets need to be comparable for discrete modelling.

4 Conclusion

In this work, we demonstrated that enforcing the discreteness of raw transcriptomic data in diffusion frameworks improves generative modeling in unconditioned and conditioned settings. We introduced DCM, a discrete diffusion model that achieves new state-of-the-art performance in unconditioned generation on the Dentate Gyrus dataset, indicating the model has formed better representations of global and granular transcriptomic patterns. On the Replogle dataset, DCM similarly improves W_2, confirming stronger cell population-level alignment, while performance on MMD^2RBF is more variable. This suggests that while discrete diffusion robustly captures global transcriptomic structure across datasets, fine-grained dependency modeling may depend on conditioning complexity, highlighting an important direction for future work.

Beyond the empirical improvements, this work demonstrates that generative models gain representational power when their state space matches the discrete, sparse structure of the biological measurements they aim to model. While applied here with single-cell transcriptomics data, this principle also extends to other count-based molecular assays that will enable more faithful virtual cells.

References

Britt Adamson, Thomas M Norman, Marco Jost, Min Y Cho, James K Nuñez, Yuwen Chen, Jacqueline E Villalta, Luke A Gilbert, Max A Horlbeck, Marco Y Hein, et al. A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882, 2016.

Abhinav K Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, et al. Predicting cellular responses to perturbation across diverse contexts with state. *BioRxiv*, pp. 2025–06, 2025.

Michael Bereket and Theofanis Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. *Advances in Neural Information Processing Systems*, 36:1–12, 2023.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and A. Doucet. A continuous time framework for discrete denoising models. *ArXiv*, abs/2205.14987, 2022. URL https://api.semanticscholar.org/CorpusID:249192370.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7): 1853–1866, 2016.

Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

W Tecumseh Fitch. Information and the single cell. *Current Opinion in Neurobiology*, 71:150–157, 2021.

Shreshth Gandhi, Farnoosh Javadi, Valentine Svensson, Umair Khan, Matthew G Jones, John Yu, Daniele Merico, Hani Goodarzi, and Nima Alidoust. Tahoe-x1: Scaling perturbation-trained single-cell foundation models to 3 billion parameters. *bioRxiv*, pp. 2025–10, 2025.

Nate Gruver, Samuel Stanton, Nathan C Frey, Tim G. J. Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew Gordon Wilson. Protein design with guided discrete diffusion. *ArXiv*, abs/2305.20009, 2023. URL `https://api.semanticscholar.org/CorpusID:258987335`.

Gunsagar S Gulati, Shaheen S Sikandar, Daniel J Wesche, Anoop Manjunath, Anjan Bharadwaj, Mark J Berger, Francisco Ilagan, Angera H Kuo, Robert W Hsieh, Shang Cai, et al. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*, 367(6476):405–411, 2020.

Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology*, 20(1):296, 2019.

Hannah Hochgerner, Amit Zeisel, Peter Lönnerberg, and Sten Linnarsson. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell rna sequencing. *Nature neuroscience*, 21(2):290–299, 2018.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. *arXiv preprint arXiv:2502.06768*, 2025.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Romain Lopez, Jeffrey Regier, Michael Cole, Michael Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15, 12 2018a. doi: 10.1038/s41592-018-0229-2.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018b.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):e11517, 2023.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning*, 2023. URL https://api.semanticscholar.org/CorpusID:264451832.

Erpai Luo, Minsheng Hao, Lei Wei, and Xuegong Zhang. scdiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics*, 40, 2024. URL https://api.semanticscholar.org/CorpusID:266844972.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Chenlin Meng, Kristy Choi, Jiaming Song, and Stefano Ermon. Concrete score matching: Generalized score matching for discrete data. *Advances in Neural Information Processing Systems*, 35: 34532–34545, 2022.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *International Conference on Learning Representations (ICLR)*, 2025.

Giovanni Palla, Sudarshan Babu, Payam Dibaeinia, James D Pearce, Donghui Li, Aly A. Khan, Theofanis Karaletsos, and Jakub M. Tomczak. Scalable single-cell gene expression generation with latent diffusion models. *ArXiv*, abs/2511.02986, 2025. URL https://api.semanticscholar.org/CorpusID:282758604.

Alessandro Palma, Till Richter, Hanyi Zhang, Manuel Lubetzki, Alexander Tong, Andrea Dittadi, and Fabian J Theis. Multi-modal and multi-attribute generation of single cells with cfgen. In *International Conference on Learning Representations*, 2024. URL https://api.semanticscholar.org/CorpusID:271218151.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *ArXiv*, abs/2406.07524, 2024. URL https://api.semanticscholar.org/CorpusID:270380319.

Sophia Tang, Yinuo Zhang, and Pranam Chatterjee. Peptune: De novo generation of therapeutic peptides with multi-objective-guided discrete diffusion. *ArXiv*, 2024. URL https://api.semanticscholar.org/CorpusID:274992140.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Tao Zeng and Hao Dai. Single-cell rna sequencing-based computational analysis to describe disease heterogeneity. *Frontiers in Genetics*, 10:629, 2019.

Appendix

A DATASETS USED

Table 3: Summary of datasets used in the experiments.

Experiment	Dataset name	No. of cells	No. of genes	No. of cell types/lines
1	Dentate gyrus	18,213	17,002	14 cell types
2	Replogle-Nadig	624,158	2,000 (HVGs)	4 cell lines

B TRAINING DETAILS

Table 4: Hyperparameter values of DCMs considered in this paper.

DCM - score entropy matching	
Denoising Transformer	
Number of Blocks	8
Number of Heads	8
Embedding size	256
Normalization	LayerNorm
Adaptive Normalization	True
Hyperparams	
σ	$1e^{-4}$

C VOCABULARY SIZE AND COMPUTATIONAL COMPLEXITY

A potential concern with defining a discrete vocabulary $\{0, 1, \ldots, V_{\max}\}$ over raw expression counts is that $V_{\max}$ could become large for high-sequencing-depth datasets, posing computational and memory bottlenecks. However, this concern does not arise in our setting. Since we employ absorbing diffusion, the transition probabilities and score computations have closed-form solutions that do not require materializing any structure of size $V_{\max}$; all required quantities are computed analytically on the fly, yielding $\mathcal{O}(1)$ complexity with respect to $V_{\max}$ for the diffusion calculations. The only component that scales with $V_{\max}$ is the transformer's embedding and output projection layer—a standard linear layer that is not unique to our model. Discrete diffusion language models routinely operate over vocabularies of $50{,}000$ tokens or more (Sahoo et al., 2024) and have been scaled to billions of parameters without this being a practical bottleneck. The values of $V_{\max}$ typical of single-cell RNA-seq datasets fall well within this regime.

D CODE

The implementation of the methods discussed in this paper is available online at [`https://github.com/sanjukta7/aivc-dcm`].